Attention, Perception, & Psychophysics
2010, 72 (7), 1825-1838
doi:10.3758/APP.72.7.1825

The visual system discounts emotional deviants when extracting average expression

Jason Haberman
University of California, Davis, California

and

David Whitney
University of California, Davis, California
and University of California, Berkeley, California

There has been a recent surge in the study of *ensemble coding*, the idea that the visual system represents a set of similar items using summary statistics (Alvarez & Oliva, 2008; Ariely, 2001; Chong & Treisman, 2003; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). We previously demonstrated that this ability extends to faces and thus requires a high level of object processing (Haberman & Whitney, 2007, 2009). Recent debate has centered on the nature of the summary representation of size (e.g., Myczek & Simons, 2008) and whether the perceived average simply reflects the sampling of a very small subset of the items in a set. In the present study, we explored this further in the context of faces, asking observers to judge the average expressions of sets of faces containing emotional outliers. Our results suggest that the visual system implicitly and unintentionally discounts the emotional outliers, thereby computing a summary representation that encompasses the vast majority of the information present. Additional computational modeling and behavioral results reveal that an intentional cognitive sampling strategy does not accurately capture observer performance. Observers derive precise ensemble information given a 250-msec exposure, suggesting a rapid and flexible system not bound by the limits of serial attention.

At any given moment, we are actively identifying one or very few items (Luck & Vogel, 1997; Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1998; Simons, Nevarez, & Boot, 2005). However, this counters our rich perceptual experience. Either our intuition is an illusion (often referred to as the *grand illusion*; Noë, Pessoa, & Thompson, 2000) or we perceive more information than has been revealed by previous studies. There has been a recent surge in the study of *ensemble coding*, the visual system's ability and natural tendency to represent sets of similar items using summary statistics (Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2007, 2009; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). For example, averaging has been established for low-level features and textures, such as orientation (Dakin & Watt, 1997; Parkes et al., 2001), direction and speed of motion (Watamaniuk, 1993; Watamaniuk & Duchon, 1992), position (Alvarez & Oliva, 2008; Morgan & Glennerster, 1991), shadow orientation (Koenderink, van Doorn, & Pont, 2004), size information (Ariely, 2001; Chong & Treisman, 2003), and even facial expression/identity (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007, 2009; Sweeny, Grabowecky, Paller, & Suzuki, 2009). Although it is generally agreed that the visual system extracts summary statistical properties about most low-level features (e.g., position, orientation, motion, etc.), it is unclear whether this is the case for information about sizes of averages (Myczek & Simons, 2008).

In a series of elegant simulations, Myczek and Simons (2008) found that many of the results purporting average size perception could be explained by our current understanding of working memory (i.e., sampling just a few items from the set), negating the necessity of a putative mechanism dedicated to average size representation (although this is debated; Ariely, 2008; Chong, Joo, Emmanouil, & Treisman, 2008; Simons & Myczek, 2008). Although Myczek and Simons's work was limited strictly to average size perception, it is important to consider that such a cognitive sampling strategy could extend to other visual domains (e.g., faces); perhaps summary statistical representations are automatically computed, prior to a stage of selection, only for very low-level visual features, such as position, motion, color, lightness, and orientation.

We have previously shown that observers perceive the average expression in a crowd of faces with great precision and reliability (Haberman & Whitney, 2007, 2009). Observers derive this mean representation despite lack-

J. Haberman, haberman@wjh.harvard.edu

ing information about the set's constituents. However, an important question is whether the perception of average facial expression is actually the result of a dedicated, automatic, summary statistical process (e.g., linear pooling, as in the orientation domain; Parkes et al., 2001). It may be the case that the perception of average facial expression results from the cognitive sampling of one or two face(s) from the entire set, rather than from an explicit averaging mechanism.

It is important to note that, in all visual domains—whether texture perception, global motion perception, or average expression perception—subsampling some number or percentage of stimuli over space or time will eventually adequately explain performance (e.g., Morgan, Chubb, & Solomon, 2008). Thus, we distinguish between *automatic, implicit, unintentional subsampling*, which is ubiquitous even in "low-level" texture perception, and *cognitive subsampling*, in which intentionally collecting only one or two items is sufficient to match averaging performance. Our goal was to test whether ensemble perception occurs automatically and unintentionally and is not bound by the limits of a cognitive sampling strategy.

Given the complexity of emotional face processing, one might expect that perceiving average facial expression should rely on a serial sampling of very few face images (i.e., cognitive subsampling). Indeed, emotion recognition is processed with relative sluggishness; faces typically do not exhibit popout effects (Brown, Huey, & Findlay, 1997; Kuehn & Jolicœur, 1994; Nothdurft, 1993), even from among other faces that possess highly salient differences. Searching for an emotional face tends to be a deliberate, serial process (although see Hansen & Hansen, 1988). Ensemble coding of faces—the perceptual averaging of emotional expression (Haberman & Whitney, 2007, 2009)—might therefore be expected to be a serial, deliberate process as well.

Here, we show that, on the contrary, summary information about groups of faces is derived very quickly, is sensitive to overall statistics of crowds of faces, and is not driven by cognitive subsampling of 1–3 items. We measured sensitivity to summary statistics (i.e., average expression in groups of faces) using a method-of-adjustment (MOA) technique. Observers adjusted a test face to match the perceived mean expression of a preceding set of faces that contained emotional outliers. The emotional outliers introduced additional variance into the set, which made summary representation more difficult (Morgan et al., 2008). We examined whether observers would compensate for the increased variance by preferentially representing the *local mean* (the mean of the set, excluding the outliers) over the *global mean* (the mean of all of the items in the set). Through three behavioral experiments, along with Monte Carlo simulations, we show that observers more precisely represented the local mean expression of a 12-item set after only a 250-msec exposure. A cognitive subsampling strategy cannot adequately account for the speed and precision of this effect. These experiments provide evidence that, under some circumstances, the visual system coarsely codes summary statistics about the expressions of faces in crowds, all while discounting deviant information.

EXPERIMENT 1A

In the first experiment, we used an MOA technique to assess the precision with which the observers would represent the mean emotion of a set of faces. Observers adjusted the expression of a test face to match the perceived mean of the previously displayed set of faces. Unlike in previous research, however, these sets contained emotional outliers (i.e., faces that substantially deviated from the overall set mean). The introduction of emotional outliers addresses a particular issue of interest: Do observers incorporate the outliers in their assessment of the overall set mean, or do they disregard the outliers? Outlier facial expressions would tend to disrupt a serial sampling strategy (Myczek & Simons, 2008) more than a global, parallel averaging process (e.g., linear pooling, as in the orientation domain; Parkes et al., 2001) because including outliers in one's sample would heavily distort or shift the mean representation (particularly if observers sampled only a couple of items).

The MOA technique offers an alternative measure of summary statistical precision, as described below. It is important to note that we pushed the limits of statistical face representation abilities by presenting sets of 12 faces for 250 msec.

Method

Participants. Five individuals (3 female, mean age = 23.6 years) affiliated with the University of California at Davis, participated. Informed consent was obtained for all of the volunteers, who were compensated for their time and had normal or corrected-to-normal vision. All of the research was approved by the university's Institutional Review Board.

Stimuli. We created three sets of 50 faces by linearly interpolating (Morph 2.5, 1998) between 2 emotionally extreme faces of the same person, taken from the Ekman gallery (Ekman & Friesen, 1976). To create the range of morphs, multiple facial features (e.g., corners of the mouth, bridge of the nose, center of the eye, etc.) were matched between the emotionally extreme faces. The software then linearly morphed between the start- and endpoints specified and outputted 50 image files. The stimulus sets ranged from happy to sad, sad to angry, and angry to happy. The amalgam of 150 faces formed the stimulus set, a virtual circle of emotions that was functionally infinite (see Figure 1).

Morphed faces were nominally separated by emotional units (e.g., Face 2 was one emotional unit sadder than Face 1). The label *emotional unit* is arbitrary, and we do not mean to imply that every emotional unit corresponds to a categorically distinct emotion. Although emotion representation is thought to unfold nonlinearly in emotion space (Russell, 1980), previous testing revealed that our stimulus set is psychophysically linear (i.e., all morphs were equally discriminable; Haberman, Harp, & Whitney, 2009). Face images were grayscale (98% maximum Michelson contrast) and occupied 3.04º × 4.34º of visual angle. The set of 12 faces on the screen occupied 12.16º × 13.02º of visual angle. Faces were presented in a 3 × 4 grid (Figure 1A). The background relative to the average face had a maximum Michelson contrast of 29%.

In previous studies, ensemble face perception was investigated using sets of faces that had a uniform distribution of emotional valences (Haberman & Whitney, 2007, 2009). Similar to those in

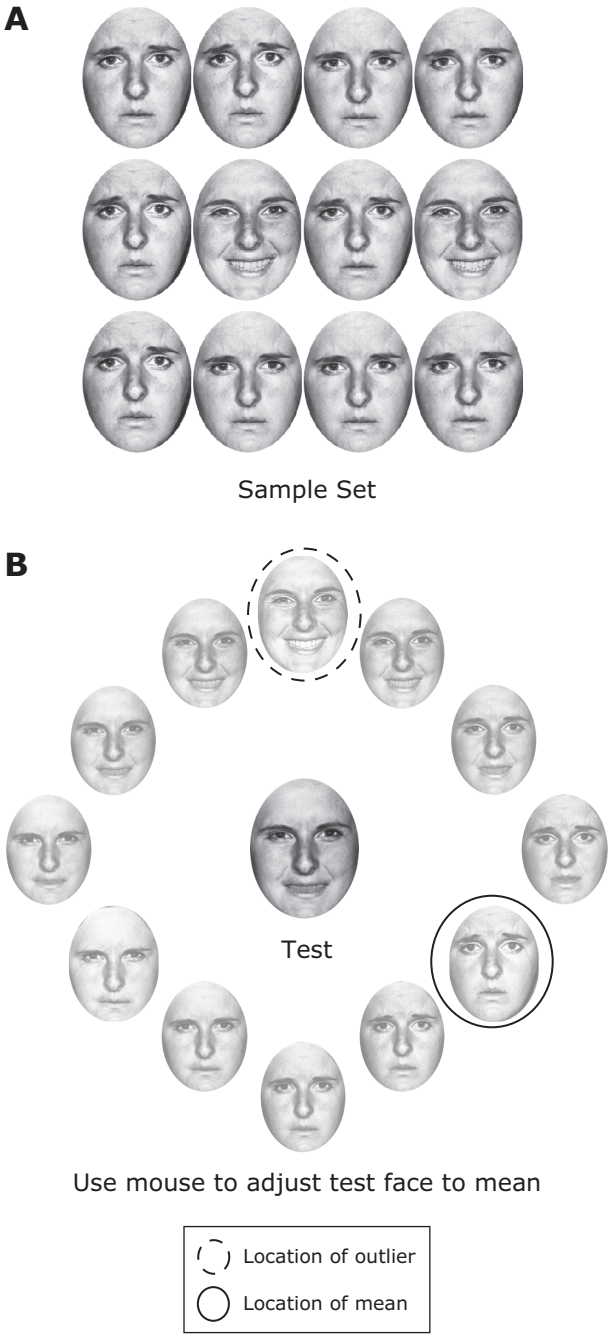

Figure 1. (A) Sample trial from Experiment 1. Observers viewed sets of 12 faces for 250 msec. Each set contained 2 emotional deviants. (B) Circle of emotions used in the experiments. A random face along this continuum was displayed during the test phase, and observers used the mouse to adjust the test face to match the emotional mean of the previously displayed set. The solid circle represents the set mean (local), and the dotted circle indicates the location of the deviants. Note that this is a sparse representation of the stimulus set.

previous experiments, the sets in this experiment initially contained three instances of four emotional expressions. Faces were ±3 and ±9 emotional units around a randomly selected set mean, and all set members were distinguishable from each other. The difference in this experiment was that 2 of the members, selected randomly, were replaced with emotional outliers; sets therefore contained 10 faces ±3 and ±9 emotional units around a randomly selected mean and 2 identical outlier faces whose expression was ±60 units away from the initial set mean (see Figure 1). This skewed the distribution and mean expression away from a uniform distribution and increased overall set variance. Note that there was a local mean expression corresponding to the average expression of the 10 faces that were within ±9 emotional units of each other (excluding the outliers). There was also a global mean corresponding to the average expression of all 12 faces on the screen (i.e., incorporating the outlier faces that were ±60 emotional units away from the local mean).

Procedure. Observers saw the set of 12 items for 250 msec, followed by a single test face. The initial expression of the test face was random. Using the mouse, observers adjusted the test face to match the perceived average expression of the preceding set. The adjustment task allowed observers to cycle through the morph circle (Figure 1) and choose any 1 face from the set of 150. Observers pressed the left mouse button to indicate their choice, and the next trial began 500 msec after the buttonpress.

Each run had 200 trials, and observers performed four to six runs (800–1,200 total trials).

Most previous experiments exploring summary statistics (Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2007) incorporated some form of a two-alternative forced choice (2AFC) paradigm in which chance performance was 50% correct. However, chance performance on an MOA task is defined as 1 divided by the number of stimuli (1/150 in our experiment). Rather than categorizing each response as *correct* or *incorrect*, however, MOA allows us to derive how far observers were from the actual set mean on every trial. In other words, we can plot observers' complete error distributions.

Results and Discussion

On each trial, there was a local mean expression and a global mean expression. The local mean was the average emotion of all set members, excluding the two outliers. The global mean was the average emotion of all set members. Figure 2 shows the error distribution around the local and global means—that is, the difference between the observer's selected test face on each trial and local and global means, respectively. A Von Mises distribution (a circular normal distribution) was fit to the error distribution. Unlike in a Gaussian distribution, the area under a Von Mises curve must integrate to 1. Since our stimuli formed an emotional circle, the Von Mises was the appropriate distribution to use. It was formalized as

$$\left\{\frac{\exp[k * \cos(x-a)]}{[2\pi * besseli(0,\ k)]}\right\},$$

where a was the location (i.e., where along the circle the points cluster) and k was the concentration (i.e., inversely related to *SD*, so the larger the number, the more concentrated the distribution). We assessed the precision of mean representation by measuring the *SD* (converted from k) of the Von Mises distribution; the smaller the *SD* of the curve, the more precise the mean representation. Figure 2 shows that all of the observers had a smaller *SD* for local mean than for global mean, suggesting that the observers were more sensitive to the local mean of the set. Monte Carlo resampling revealed that the error distributions for the 5 individual observers were narrower for the local mean than for the global mean, and statistical tests

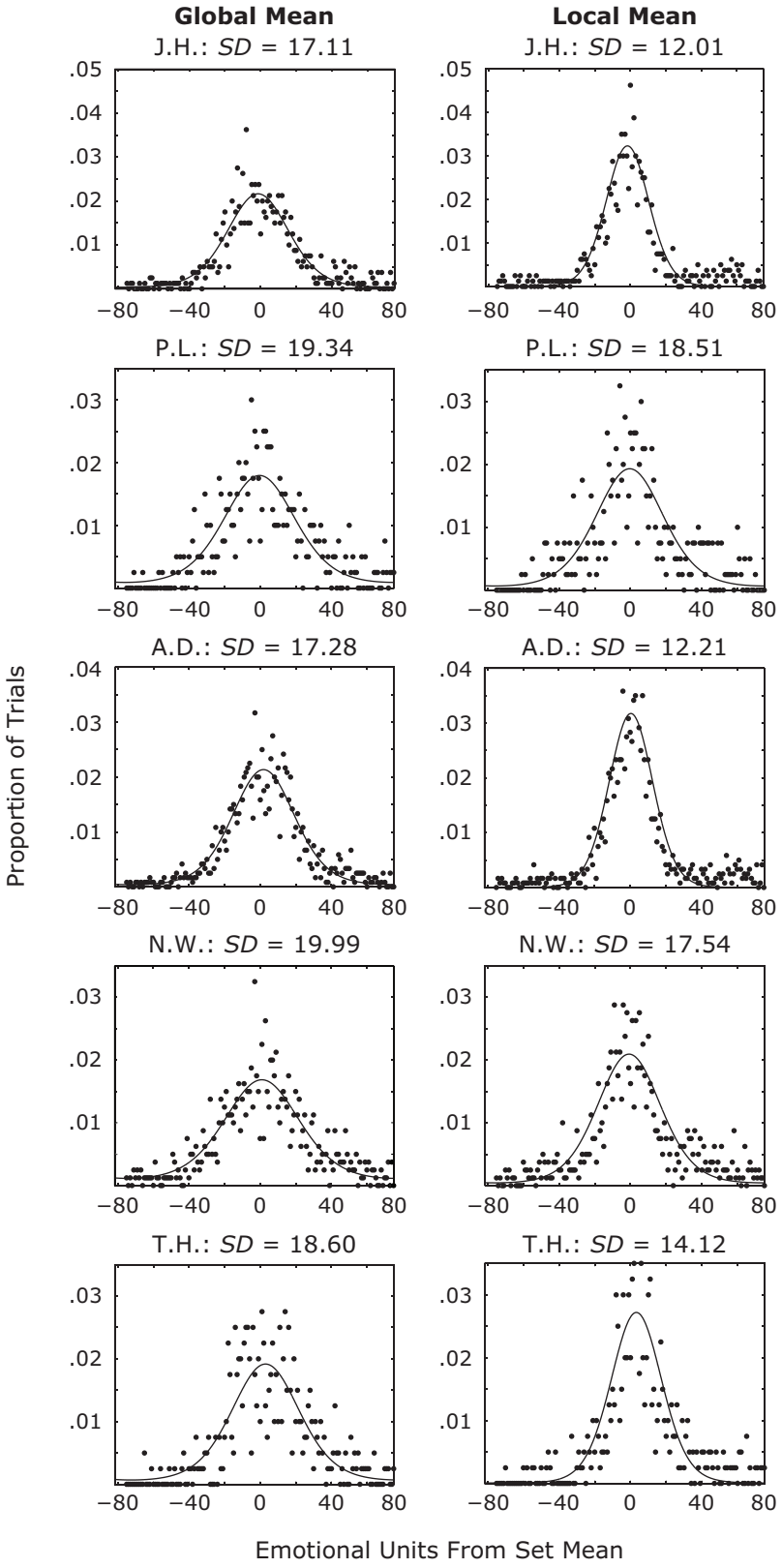

Figure 2. Experiment 1's results, including the method-of-adjustment error distributions and the proportion of responses at each possible separation between the user-selected test face and the actual set mean. For each observer, we plotted distance from the global mean (including outliers) and distance from the local mean (excluding outliers) and fit a Von Mises distribution to the data. Note the narrowing width of the curve in the local mean condition relative to the global mean (also reflected in smaller *SD*s), consistent for each observer. The axes have been converted from radians to emotional units for readability.

reached significance for 4 observers (only P.L.'s was not significant). A paired *t* test across observers also revealed a significantly smaller *SD* (i.e., better precision) for local mean than for global mean [$t(4) = 4.26$, $p = .013$]. The greater sensitivity to the local mean suggests that observers either filtered or suppressed the outlier information—that is, that which was impossible to integrate with the rest of the set.

The analysis described above is collapsed across the sign of the emotional outliers. That is, it treats outliers occurring 60 units above the mean the same as outliers occurring 60 units below the mean. The sign of the outlier may be informative, however, so we reanalyzed the data, fitting curves to trials in which observers saw only high or only low outliers (see Figure 3A). The results provide additional evidence that observers filter, at least to some ex-

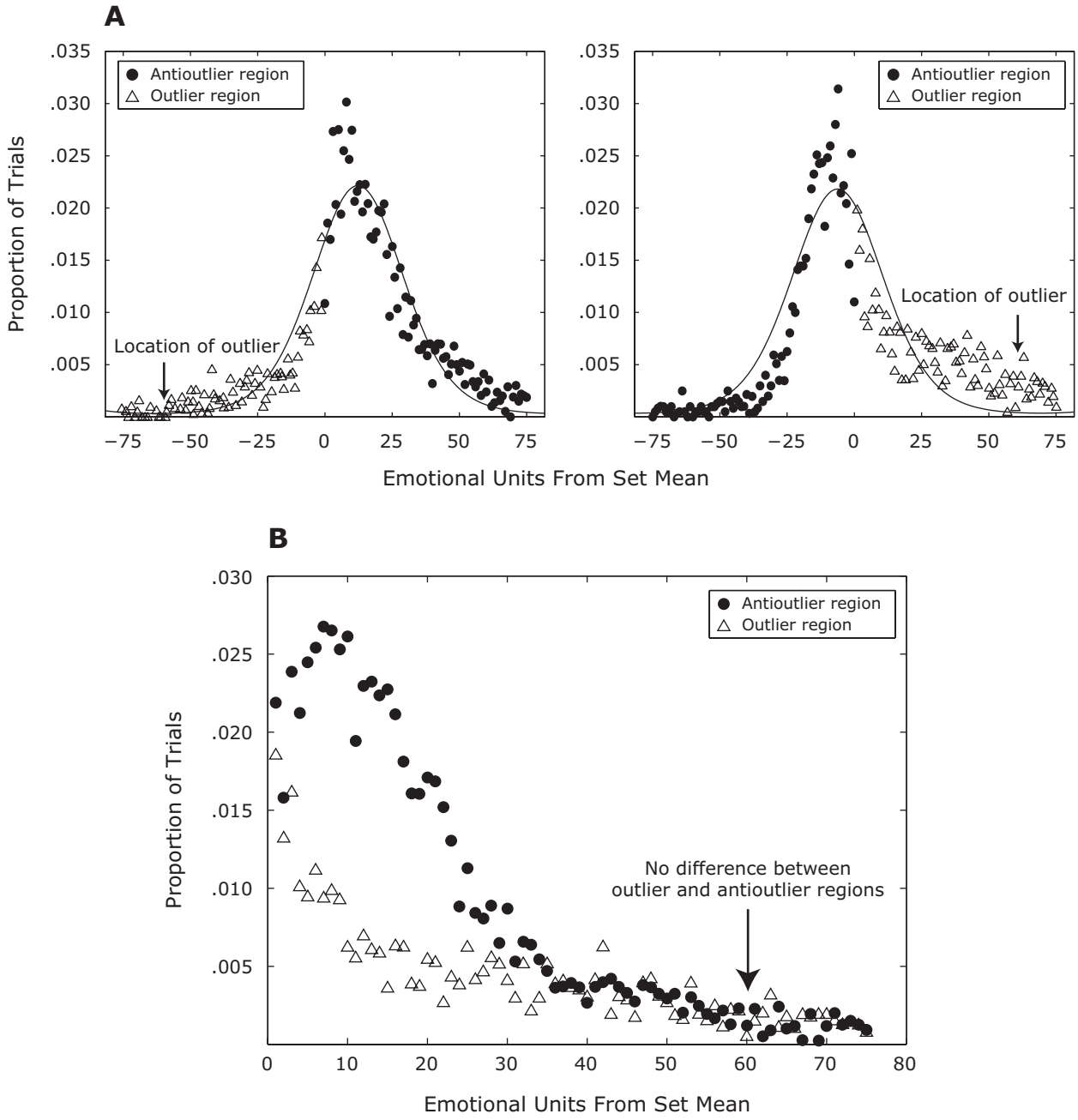

Figure 3. Analysis when the signs of the outliers are taken into account. (A) Regardless of the sign of the outliers, observers tended to ignore them when adjusting to the mean of the set, indicated by the offset of the curve from 0. (B) When flipping and averaging the data from panel A, there is no significant difference in the proportion of responses that occurred at the outlier location as compared with those that occurred at the antioutlier location (i.e., the areas surrounding 60 units from the mean).

tent, outlier information in the set. Figure 3A shows that, when the outliers are negative, adjustment to the global mean is offset in the *positive* direction. The peak of the adjustment curve corresponds closely to where the local mean occurs (i.e., the mean when outliers are excluded), approximately 10 emotional units above the global mean. The converse is also true (see Figure 3A). This trend is consistent among all observers, and paired *t* tests confirm that the absolute offset (the *a* parameter from the Von Mises equation; how far away from 0 the curve peaks) is larger for outliers in the global mean adjustment curve than for those in the local mean adjustment curve [$t(4) = 4.20, p = .014$].

One might expect that, if the analysis were broken down by the sign of the outlier, there would be a disproportionate amount of observer responses corresponding to the location of the outlier, relative to the region where there was no outlier. However, a paired *t* test examining the proportion of responses that occurred in the outlier regions with those that occurred in the antioutlier regions revealed no difference [$t(10) = 0.76, p = .31$][1] (see Figure 3B). This indicates that, after viewing a set with an outlier, observers were not more likely to pick a face near the outlier than they were to pick a face far from the outlier. This definitively rules out the possibility that observers subsampled only one item from the entire set, since, had this been the case, there would have been a greater number of responses in the outlier region than in the antioutlier region. Subsampling is further explored in Experiment 2.

This experiment demonstrated that observers have greater sensitivity to the local mean of the set than to the global mean, and therefore seem to discount the emotional outliers. The value of a summary statistic, such as average texture, orientation, motion, or facial expression, is only useful insofar as it captures ensemble information in the stimulus. Outliers disrupt ensemble information by increasing set variance; therefore, discounting the deviant information might be advantageous. Outlier discounting is a computationally simple way to mitigate variance and increase the reliability of a summary statistic, such as average expression.

EXPERIMENT 1B

It is possible that, instead of implicitly filtering deviant information, observers were aware of the presence of the outliers on every trial and consciously ignored them in their estimate of the average expression. Such a strategy would indicate that the emotional outliers can be detected and may even pop out relative to the rest of the faces, which could undermine the computational efficiency with which ensemble coding is thought to operate (Alvarez & Oliva, 2009). Although some work suggests that emotionally deviant information is not available preattentively (Nothdurft, 1993), there is at least some (controversial) evidence to suggest that it is (Hansen & Hansen, 1988). Even if observers did not detect the emotional deviance in a parallel fashion, other low-level features may have distinguished the outlier faces. In a separate control experiment, we explicitly tested whether the outlier faces popped out and how well observers represented those faces. If the representation of the outliers was poor and they were no more detectable than other faces in the set, a strategy in which observers consciously or deliberately discount the outliers would be improbable.

Method

Participants. Three naive observers affiliated with the University of California at Davis (2 female, mean age = 23 years) were tested; 2 of these observers had not participated in Experiment 1A.

Stimuli and Procedure. All stimuli were the same as those described in Experiment 1A. Observers had to adjust a test face to match any deviant information (any outlier face) in the set across three intermixed conditions (see Figure 4A): (1) sets containing no outliers (i.e., catch trials), (2) sets identical to those in Experiment 1A (2 identical outliers), and (3) sets containing outliers that were highlighted by a red box (i.e., explicit popout trials). If outlier information pops out, performance in adjusting to the outlier condition (Condition 2) should be similar to performance in adjusting to the outlier in the popout condition (Condition 3). However, if observers do not have explicit access to the individual elements of the set, outlier or otherwise, then we would expect outlier adjustment (Condition 2) to be closer to performance on the catch trials (Condition 1). Observers were not told that there were catch trials. Observers completed two runs of 300 trials each (200 trials per condition, 600 trials total).

Results and Discussion

We assessed adjustment precision using the mean square error (MS_e), which reflects how far, on average, observers were from the actual outlier face: The larger the MS_e, the worse the representation. Figure 4B shows the MS_e for each condition for 3 observers[2] (confidence intervals were derived from 1,000 bootstrapped estimates). The MS_e for outlier adjustment was not significantly different from the MS_e for the no-outlier condition, suggesting that observers were near floor when trying to adjust to the outlier face. On the other hand, performance in the popout condition (i.e., red box highlighting the outliers) was significantly better than that in the outlier condition (bootstrap test, $p < .01$). This supports the idea that the deviant expressions (outliers) did not pop out. This pattern was consistent across the 3 observers.

Interestingly, performance for homogeneous discrimination (i.e., adjusting to a single test face in a known location in the set) was still significantly better than performance in the popout condition (see Figure 4B; $p < .05$). This suggests that, even when the elements of the set pop out, such popout does not confer the same performance advantage that knowing the actual target location does. This further weakens the argument that observers could have a robust representation of sampled elements, identify the deviants, and then consciously remove them from their estimate of the average expression.

Despite the observers' generally poor performance in the outlier condition, it may be possible to calculate an upper estimate of what proportion of the time the outliers could have popped out. Here, the no-outlier condition (catch trials) displayed stimuli that could never pop out, whereas the popout condition (outliers surrounded in red boxes) displayed stimuli that should pop out consistently. In order to estimate the proportion of trials in which the tar-

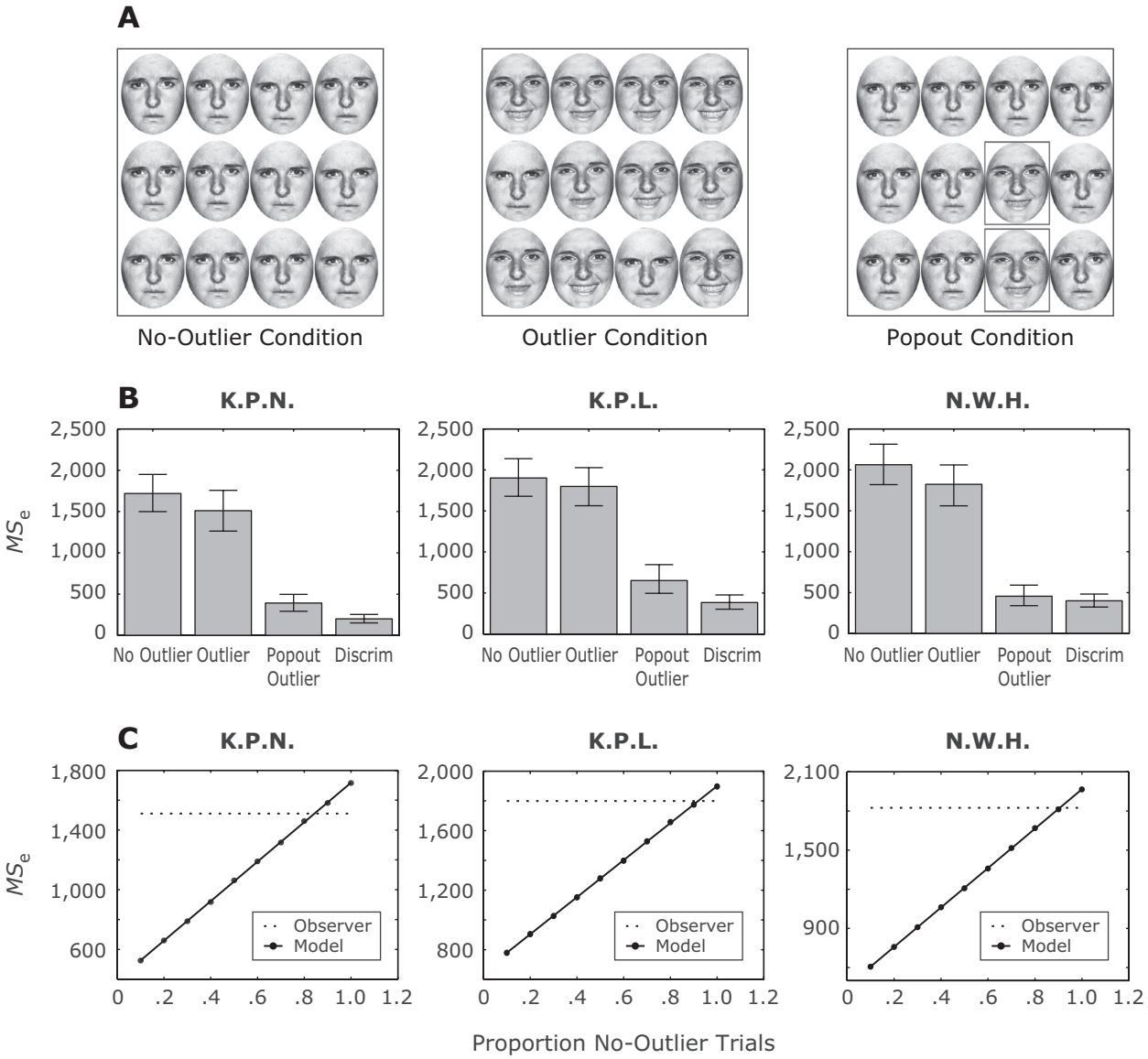

Figure 4. Stimuli and results from Experiment 1B. (A) The three conditions observers saw were sets with no outliers (catch trials), sets with outliers, and sets with red boxes around the outliers. Observers were only aware of the latter two conditions. (B) MS_e for 3 observers. Performance on outlier adjustment did not differ from performance on the catch trials; popout performance was substantially better. (C) Model used to test the proportion of outlier trials that actually popped out. Trials were randomly selected and combined in varying proportions from both the no-outlier trials and the popout trials, and, from this hybrid condition, MS_e was calculated 1,000 times. The larger the proportion of no-outlier trials, the larger the MS_e (i.e., the worse the performance). The dotted line corresponds to actual observer performance on outlier adjustment. Note that the intersection point suggests that the outliers popped out less than, on average, only ~10% of the time.

gets could have effectively popped out in the outlier condition, we modeled observer performance by bootstrapping various combinations of no-outlier and popout trials (e.g., 10% popout condition, 90% no-outlier condition). We did this 1,000 times for each of 10 possible relative proportions and calculated the MS_e for each sample (as described above) and averaged across all of the samples. As is shown in Figure 4C, the larger the proportion of no-outlier trials in the sample (fewer popout trials), the larger the MS_e. The point at which an observer's outlier condition performance intersects the modeled data provides an upper (conservative) estimate of the proportion of trials in which the outlier could have popped out (1 − intercept). It is evident that, for each observer, the outliers popped out a small fraction of the time (average ~10% of the trials, at most; see Figure 4C). This frequency of popout (even though it is not significant, on average, as revealed in Figure 4B) is not sufficient to account for the degree of outlier discounting (narrowing of the error distribution around the local mean) that is shown in Figure 2.

The evidence presented here suggests that observers had very little if any explicit knowledge of the outlier

face, which supports the conclusion that the visual system implicitly and rapidly filtered the deviant information in order to achieve a more parsimonious set representation.

EXPERIMENT 2A

Experiment 1 demonstrated a greater sensitivity to the local mean of a set than to the global mean when emotional outliers were present in the set. Additionally, it revealed that observers could not consciously discount the expressions that deviated from the rest of the set, since they were not aware of those expressions on most trials. Whereas this suggests that the visual system discounts objects that cannot be integrated into the context of the set (due to increased variance), one counterargument may be that the visual system is simply sampling from the set. An estimate of average expression may be well captured by only a couple of samples. This sampling hypothesis would be consistent with serial mechanisms of attention and was recently postulated by Myczek and Simons (2008) in response to research on average size discrimination. Their model suggests that extant data on average size do not necessitate a novel averaging mechanism, since sampling and averaging only a couple of set items proved sufficient to explain perception of average size. Despite matching human performance for many of the experiments reported in previous studies (Ariely, 2001; Chong & Treisman, 2003, 2005), their model only addresses average size. They did not test their model against sets of faces, or, more important, against sets containing outliers. However, the possibility of subsampling remains an important consideration for research on average expression.

In the present study, we created a model similar to the Myczek and Simons (2008) model that calculated a mean emotion from the sampling of 1–10 items from a set containing outliers. We predicted that outliers would pose difficulties for a sampling model, since the model simply averaged the information it gathered. Any outlier included in the sample would pull the model's calculated mean away from the actual mean. The purpose of this modeling procedure was to determine whether sampling a small subset would be sufficient to predict human performance on expression averaging.

Method

Procedure. This modeling was an attempt to visualize bootstrapped performance of a realistic (noisy) observer who averages some number of samples. Our model sampled 1–10 faces from the set of 12 (which contained outliers) on every trial, similar to the model of Myczek and Simons (2008). However, before averaging the sampled faces as an ideal observer would (e.g., if Face 10 and Face 16 were sampled, the ideal observer average would be Face 13), each sample selected by the model was perturbed by independent noise. The noise distribution was derived from the discrimination performance of each observer and was implemented to more accurately reflect natural perceptual and motor noise. We assessed this observer discrimination ability using the same MOA procedure described in Experiment 1, but for single face images. Observers simply adjusted the test face to match a single prespecified face in the set (rather than the mean). We then used the error distribution from this single-face adjustment task (i.e., how far observers were from the single expression) to perturb each face that the model had sampled from the set of 12. The sampled faces, perturbed by noise, were then averaged to calculate the modeled mean on a single trial. The difference between the modeled mean and the true mean of the set gives an error score. This was repeated 8,000 times in a Monte Carlo simulation to generate a complete error distribution. Separate Monte Carlo simulations were run to model an observer sampling 1–10 faces from the set of 12.

The Monte Carlo-modeled error distributions were compared with the actual observer error distributions (i.e., the error distributions were calculated relative to the local mean, since observers were more sensitive to the local mean; Experiment 1A). In this way, we could determine whether the rote sampling of any number of items could predict observer performance from Experiment 1. It is tempting to use this model as an index of the amount of information necessary for matching the ensemble coding ability seen in observers. We caution against this, however, since it is conceivable that, if observers were not samplers, no sampling condition in the model would adequately explain observer performance. Although we do not use this model to characterize the amount of information observers derive on a given trial, these results are critical to addressing the legitimate concerns raised by Myczek and Simons's (2008) model, with regard to average size.

Note that the modeling is conservative, since the averaging procedure functions as an ideal observer, perfectly averaging noisy individual face representations. Although a late-noise-averaging component seems essential (Parkes et al., 2001), it is difficult to know exactly the nature of that noise. Thus, the model in its current form is most useful for the purposes of visualization and is conservative.

Results and Discussion

Figure 5 shows the error distributions for the models that sampled 1–10 faces, overlaid with the actual error distributions, averaged across the group of observers. Each modeled curve represents expected performance of an observer who perfectly averages *N* noisy samples from the set. Visual inspection of Figure 5 reveals that the model that sampled 1 face did not adequately simulate observer performance. The sampling-1-face model occasionally sampled one of the emotional outliers, producing large bumps located far from the mean expression that observers' data did not exhibit (Figure 5). This confirms the analysis in Figure 3B—that is, that observers were not more likely to choose the outlier than the antioutlier in the set—showing that observers did not base decisions on a single sampled face.

As the model takes more samples, obvious differences between simulated and observer performance remain. One may wonder why, as the model samples more and more items, it does not approach a perfect mean representation of the set. The outlier that the model samples biases the mean representation, which explains the characteristic bumps in nearly all of the sampling conditions. As the number of samples increases, the bumps steadily move closer to the mean of the set, since the impact of the outliers is somewhat mitigated by the multiple nonoutlier samples. However, the model's strength in sampling is also its undoing: As the model samples more items, the probability of encountering an outlier approaches 100%, and the presence of that outlier will always shift the mean representation away from the actual mean (in either a positive or negative direction). Thus, the model does not effectively capture observer behavior, leading us to conclude that observers do not engage in a rote sampling strategy.

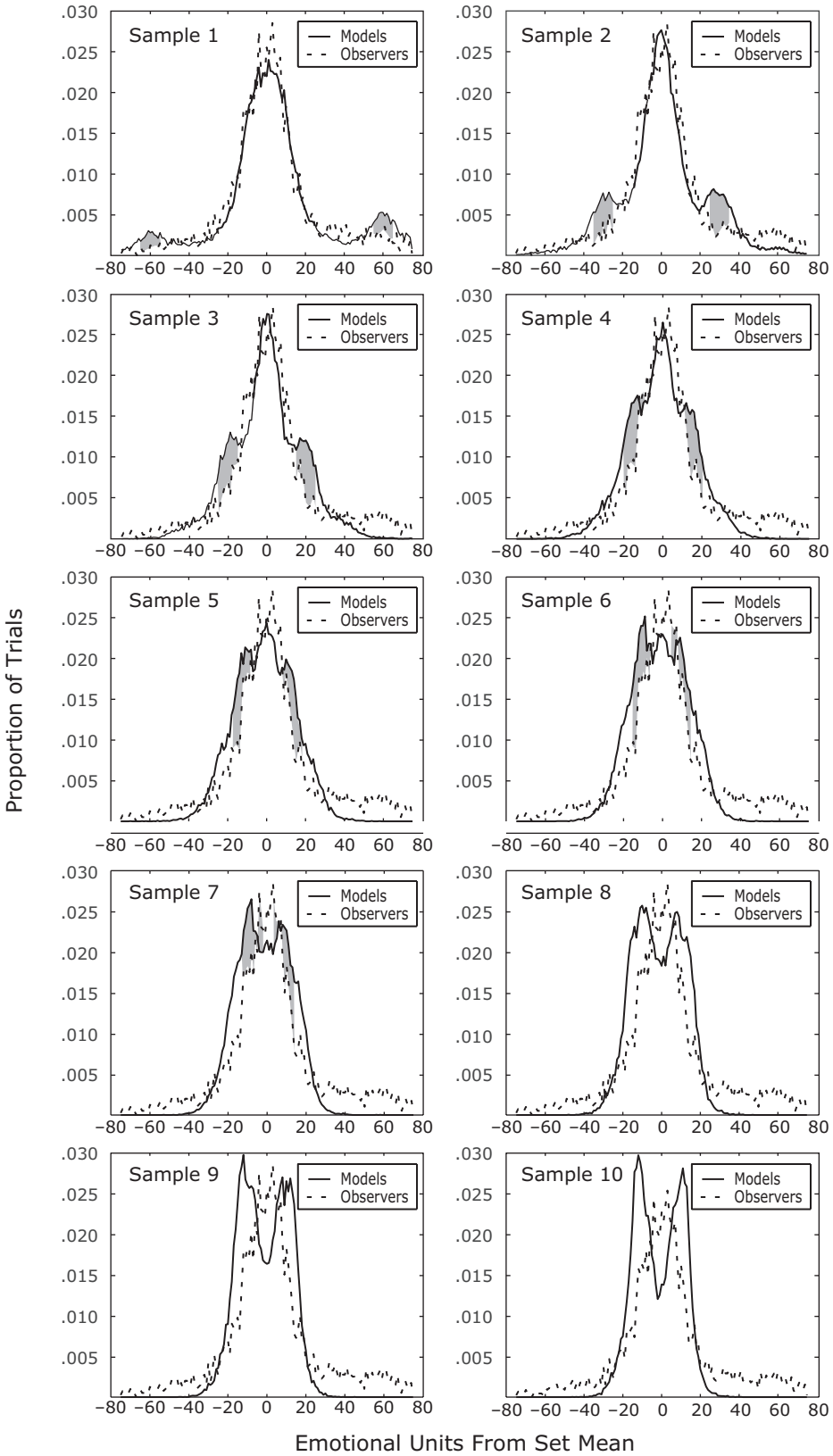

Figure 5. Experiment 2A's results, including a comparison of observer performance and model performance when averaging 1–10 samples. When taking fewer samples from the set, the model's assessment of the mean (solid line) suffers because of the outliers. Observer performance is not as adversely affected by the outliers. The gray areas highlight the bins of comparison between the model and the observer. Note that, when more than seven samples are taken, the difference between model and observer is no longer significant.

It was clear from the model's performance when sampling one face that a Von Mises distribution would not be an appropriate fit to the error distribution and is therefore not considered further as a viable assessment of the model's performance. Instead, we used a priori-defined bins to formally compare the model with the observer in each of the sampling conditions. The center of each bin was defined by the expected mean when the model encountered an outlier given N samples. For example, in the sample-three-items condition, the predicted mean representation when the model samples an outlier is ± 20 units (the outlier is ± 60, and the other two samples will, on average, be 0). This is indeed the case, as shown in the Sample 3 panel of Figure 5: There are bumps in the model error distribution at ± 20 units. The size of the bin was determined by the width at three fourths of the maximum of the group homogeneous adjustment curve (i.e., adjusted to a single known face in the set), which was approximately 11 emotional units. We used Bonferroni-corrected paired t tests to assess the proportion of trials in the model to the proportion of trials for the observers in those specified regions (shaded in each panel). Given 10 sampling conditions, our alpha level was set at $p = .005$. For Sampling Conditions 1–7, observers had significantly fewer responses in those regions than did the model, supporting the contention that the outlier had more of a differential impact on the model than on the observers. Of course, this does not imply that observers derived seven items' worth of information; it merely contrasts observers' performance with that of the model. Observers are not rote samplers. Unlike the model, they implicitly discount the deviant information in favor of the local mean.

Although the model itself does not provide an index of the amount of information observers derive from the set, such a measure would be valuable. We addressed this in Experiment 2B, where we had observers perform the behavioral equivalent of the model.

EXPERIMENT 2B

The purpose of Experiment 2B was to provide a behaviorally relevant index of the amount of information observers could accumulate about the average expression in a set of faces. To accomplish this, we assessed how well observers could adjust to the average of the set of faces when only a subset of the faces was visible.

Method

Stimuli and Procedure. The stimuli were identical to those in the first experiment, except that 8, 9, 10, or 11 of the faces in the set were randomly removed from the display. That is, only 1, 2, 3, or 4 faces from the set were visible on each trial. The task was also identical to that in the first experiment: Observers were instructed to adjust the test face to match the average expression in the set of 1–4 visible faces—a procedure somewhat similar to that used in an experiment conducted by Chong et al. (2008). The visible faces were always presented in the same positions on the grid, closest to the center of the screen, while all other faces were invisible. As in Experiment 1, observers had only 250 msec to view the face(s) and unlimited time to adjust the test face to the perceived mean of the set. Four observers (all from Experiment 1) performed five runs of 200 trials each, giving a total of 1,000 trials (250 trials for each sampling condition, 1–4) per participant.

Results and Discussion

We assessed the precision of mean representation by calculating the MS_e of the adjustment error distribution for each sampling condition (see Figure 6). Similar to the SD of the Von Mises distribution (but without the necessary assumptions), MS_e reflects how far, on average, observers were from the actual set mean: The lower the MS_e, the more accurate the mean representation. The results suggest a nonlinear pattern of mean discrimination performance, where viewing additional samples (2 or 3) actually hurt performance up to a point, after which mean discrimination improved. Critically, performance when viewing the whole set of 12 items (data from Experiment 1) was best, suggesting that a simple sampling strategy (at least for 4 or fewer items) did not account for observer performance. There was a significant difference among the sampling conditions [Friedman test: $\chi^2 = 10.2(4)$, $p = .037$]. Pairwise Wilcoxon signed rank comparisons revealed that performance for Sampling Conditions 1–3 was significantly different from performance for the whole set of 12; that is, every observer had better mean representation when presented with sets of 12 faces ($z = -1.83$, $p = .034$, one-tailed). Only Sampling Condition 4 was not significantly different from 12.

One may wonder why the data in Figure 6 are nonlinear. The reason is probabilistic and mundane. Observers were asked to adjust the test face to the mean of the entire set, given limited information. When a single face from the set was viewed, the probability of that sample being an outlier was roughly 17%. The rest of the time, the sample would be relatively close to the mean, yielding reasonable

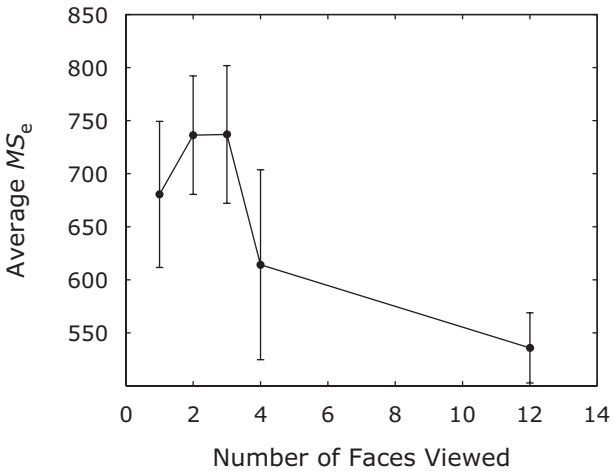

Figure 6. Experiment 2B's results, including observer performance when viewing 1–4 faces from the set, as compared with performance when viewing all 12 items. After an initial increase in MS_e (decreased mean discrimination performance), there is a significant decrease in MS_e as more information becomes available to observers. Observers adjusted the test face to the average facial expression best when all 12 face images in the set were visible.

Table 1
Probability of Sampling an Outlier

N Samples	Exactly One Outlier	Exactly Two Outliers
1	.17	N/A
2	.30	.02
3	.41	.05
4	.48	.09
5	.53	.15
6	.55	.23
7	.53	.32
8	.48	.42
9	.41	.55
10	.30	.68

mean discrimination performance (although not as good as when the whole set was viewed). However, when two samples from the set were viewed, the probability of encountering an outlier more than doubled (to around 34%; see Table 1 for the probabilities of encountering an outlier given *N* samples), further compromising local mean representation. Once the number of samples was large enough (e.g., four), the negative impact of the outliers was mitigated.

These results suggest that, in order to match observer performance when adjusting to the mean of a set of 12 faces, observers must collect at least 4 faces' worth of information. In contrast to the model described above, this task actually provides an index of information content. That is an impressive amount of information, given that observers had only 250 msec to view a set, and considering that observers were not capable of recalling the specific faces that composed the set (Haberman & Whitney, 2007). Although some iconic representation may have persisted beyond the 250-msec display (Sperling, 1960), our previous work showed that observers retained no information about the set constituents, even when there were only 4 faces on the screen displayed for 2,000 msec (Haberman & Whitney, 2007). This suggests that iconic memory confers little benefit in extracting additional information about the individuals. Furthermore, 4 faces is beyond the visual short-term memory (VSTM) capacity for these face stimuli (Haberman & Whitney, 2007, 2009).

These data also support the conclusions of the modeling in Experiment 2A. Although observers derive the equivalent of at least four faces, we cannot distinguish whether that corresponds to four discrete face representations or four faces' worth of information coarsely distributed across the entire set.

EXPERIMENT 3

In Experiment 1, we determined that observers were more sensitive to the local mean than to the global mean (Figure 2). However, we do not know whether this increased sensitivity reflects a complete suppression of outlier information or a partial discount. It is possible that, because the emotional outliers were so far removed from the rest of the faces, they could not be integrated into the context of the set and were thus suppressed. Alternatively, if observers coarsely represented several faces in the set, observers may have engaged a weighted averaging process, whereby deviant information was incorporated into the mean, but at a discount. We tested these alternatives by reanalyzing the adjustment data from Experiment 1, in which observers adjusted a test face to match the perceived mean of a set of 12. We applied a series of weights to the outlier faces, measuring performance (the adjustment error distribution) for each weight. The weighting scheme ranged from 0% (*perfect local mean representation*) to 100% (*perfect global mean representation*) at approximately 11% increments. As an illustration of how this method was implemented, take trial *N*, which has a local mean value of 100 and a global mean value of 110. The first weighting scheme would examine observer and model performance relative to a set mean of 101 (i.e., ~11% weighting of the outliers). The second weighting scheme would make the same comparisons relative to a set mean of 102 (~22% outlier weighting), and so on. By manipulating the weights, we simulated different levels of outlier suppression.

To assess mean discrimination performance at each weight, we calculated the MS_e (normalized to the grand mean) of the adjustment error distribution for each observer. The results of this analysis are plotted in Figure 7A, along with a fitted quadratic function (fit to all observers in Experiment 1). The U-shaped function that resulted from this analysis suggests that MS_e was reduced (i.e., an increase in mean representation precision) when some weight was given to the outliers (~55%). However, MS_e was higher if we assume that outliers were completely ignored or, conversely, were completely incorporated into the judgment of the mean expression, suggesting that the outliers were somewhat (but not entirely) discounted. Thus, some amount of filtering or discounting occurred as observers attempted to extract the mean expression.

Note that we also applied this outlier-weighting simulation to the sampling model described in Experiment 2A. For each of the model conditions (i.e., *N* faces sampled), we examined MS_e from 0% outlier weighting to 100% outlier weighting in 11% increments (Figure 7B). The pattern of data across all conditions was strikingly consistent, showing monotonically decreasing MS_e as the model increasingly weighted the outliers. This result emerged without regard to the number of sampled faces (1–10). This is in contrast to the pattern of results seen for observers, which showed the lowest average MS_e at a weighting of around 55% (Figure 7A). More intuitively, the simulations reported in Figure 7 reveal that, if participants were performing the task by sampling individual faces and averaging them (ideally or with added noise), they should have been more accurate when measured against the global mean than the local mean. This did not happen. This differential pattern of performance further discredits the idea that observers engage a cognitive sampling strategy when judging average expression in sets of faces.

Discussion

The experiments show that observers accurately perceived the average expression in a crowd of faces and did do so even when there were outlier faces present. Observ-

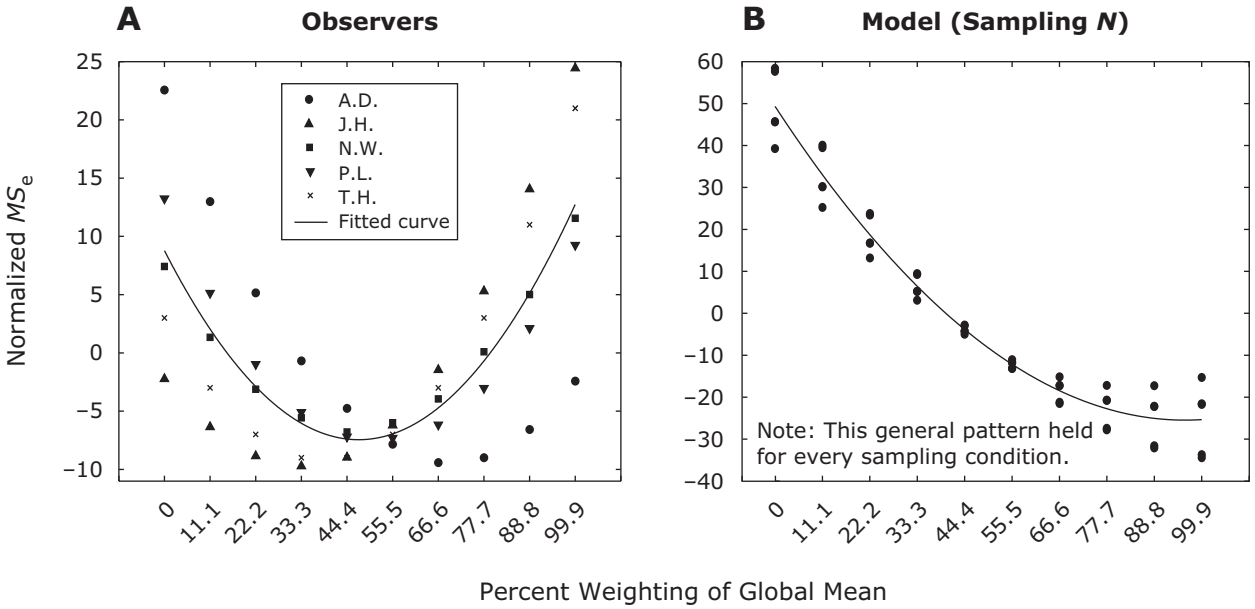

Figure 7. Experiment 3's results, including differential outlier weighting for (A) the observer data and (B) the model data. The mean representation for observers is most precise when the outliers are assigned a weight of approximately 55%; this is indicated by the lower normalized MS_e (MS_e values normalized to the grand mean). In contrast, the model's mean representation is best when it weights the outliers 100% (across all sampling conditions). This suggests that observers discount the outliers to some degree and suggests also that a sampling strategy cannot account for observer performance.

ers more precisely represented the local mean of the set, suggesting that the emotional outliers were heavily discounted. Figure 3 shows that, when the analysis is broken down by the sign of the outlier, there was little influence of the outlier, because observers unequivocally adjusted to the local mean. Critically, when the outliers were 60 units above the mean, there was no significant difference in the number of observer responses at or near that positive region, as compared with the corresponding negative region. The converse was also true. This definitively rules out the possibility that observers derived only one face from the set, since that would have resulted in a disproportionate number of responses on one side of the distribution (i.e., near the outliers).

We additionally ran Monte Carlo simulations, which revealed that observers did not employ a rote sampling strategy from the set. These experiments reinforce the findings that observers derive robust summary representations from crowds of faces (Haberman & Whitney, 2007, 2009). The results established that a sampling strategy, in which one or two items from the entire array is selected, cannot account for observer mean representation precision.

Although it is clear that observers do not behave as rote samplers, is it possible that observers still sample from the set, but in an intelligent manner, perhaps cognitively searching for and excluding the outliers? Experiment 2B suggests that observers derived at least four faces' worth of information from the set, and Experiment 1B demonstrated that observers had almost no explicit knowledge of the outliers, suggesting that the guided (cognitive or intelligent) search model is unlikely to explain the ensemble perception here.

It is true that observers might quickly sample a subset of more than four faces, implicitly discount the outliers, and then automatically compute the mean from the remaining samples. Such a process describes the data, but it characterizes the putatively automatic ensemble motion and texture perception better than it describes a cognitive or intentional search mechanism. Sampling N items, by definition, has to provide one with the solution eventually. This same principle applies to most domains of visual processing; texture and motion perception, for example, can be accounted for by calculating the effective number of sampled stimuli. Yet, there is no debate surrounding whether texture or motion perception depends on a cognitive, guided search process. That is, the possibility of subsampling, per se, should not be taken as evidence of a cognitive, serial, or attentional mechanism. We have provided evidence pointing toward a rapid process, in which deviant information is implicitly filtered or discounted in an effort to derive a precise estimate of the mean.

Although faces represent a distinctly higher level of processing (Tanaka & Farah, 1993; Young, Hellawell, & Hay, 1987), our results suggest that groups of faces are processed in a manner similar to groups of oriented lines, moving dots, textured elements, and other low-level features; sets of faces form a kind of texture, where the individual face representations are lost (Experiment 1B; see also Haberman & Whitney, 2007, 2009), but the facial group, as a gestalt, is maintained.

Our results are particularly interesting in the context of some recent work in the average size domain (de Fockert & Marchant, 2008) and may serve to further distinguish average size perception from average expression perception.

In their study, de Fockert and Marchant asked observers to locate either the largest or smallest circle in the set and then assess the average size of the set as a whole. Directed attention to the most extreme element in the set biased average size representation (e.g., locating the largest circle in the set resulted in larger and often incorrect average size representation) and may suggest that observers were subsampling. In contrast, our results showed that average expression representation was relatively unaffected by the outliers. The key difference between our results and those of de Fockert and Marchant was the attentional manipulation. Our results here and in previous studies (Haberman & Whitney, 2007, 2009) suggest that ensemble expression perception is fast, automatic, implicit, and relatively insensitive to outliers. However, we cannot conclude that attention plays no role. Indeed, recognizing any face—even a single face—may involve attention (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Wojciulik, Kanwisher, & Driver, 1998). Whether recognizing an ensemble (group) of faces requires any additional attention, however, remains an intriguing question. More generally, although coding certain types of stimuli requires some degree of attention, there may be little or no *added* cost for representing the ensemble properties of those stimuli. Ensembles may come for free, even if the stimuli on which they are based do not.

Why would the visual system selectively filter outlier expressions from the representation of the set of faces? The outliers are set deviants and tend to cross at least one emotional category. For example, it might not be perceptually meaningful to extract the absolute mean of a happy face, an angry face, and a sad face. Instead, the visual system may filter or discount information that cannot be effectively integrated into the summary statistics of the set.

There might also be a statistical basis for observers' reduced sensitivity to the global mean. Adding outliers to a set of faces increases the emotional variance of the set, and added variance might make summary statistical recognition more difficult. Other work on texture discrimination, in which observers were asked to judge which of two textures had the greater orientation variance (another type of summary statistic), showed that the task was more difficult (thresholds were higher) when orientation variance was high (Morgan et al., 2008). We suggest that, when the variance of a set is too high, the mean loses its utility. Thus, the visual system may discount the outlier information in order to maintain a more useful summary.

The mechanism by which summary statistics, such as average expression, are extracted from groups of faces remains unknown. It could be a weighted pooling (averaging) of face representations by a population of neurons, each of which has broad but overlapping tuning for expression. Such linear pooling is thought to occur in the orientation and motion domains (Parkes et al., 2001; Watamaniuk, McKee, & Grzywacz, 1994). How the outliers are discounted, however, is less clear. Either an additional stage is necessary to introduce that nonlinearity, or, perhaps more simply, the visual system might compute a median expression. The median is another reasonable summary representation and is much more robust to the adverse effects of outliers. The mode might also work; it offers a winner-take-all method of deriving a summary, which would automatically discount the deviants. One way to identify more precisely what measure of central tendency (mean, median, or mode) the visual system extracts would be to skew the distribution of the sets (Chong & Treisman, 2003). Computational modeling akin to that described in Experiment 2A would also be an effective method of addressing this question in future studies.

Regardless of which summary statistic is being represented (weighted mean, median, or some hybrid), the data here suggest a remarkably fast and flexible process. The visual system is summarizing more than four faces' worth of information in as little as 250 msec (16 Hz). This speed is impressive, given that attentional dwell time is estimated at between 200 and 500 msec (Duncan, Ward, & Shapiro, 1994; Wolfe, 2003), serially searching through sets of faces takes between 70 and 150 msec per face (Nothdurft, 1993; Tong & Nakayama, 1999), and observers have no VSTM of any of the specific faces that compose the sets (Haberman et al., 2009; Haberman & Whitney, 2007, 2009). Although visual attention may process several objects simultaneously (Wolfe, 2003), our results demonstrate that what emerges is not a representation of any individual face but an entirely novel summary percept.

AUTHOR NOTE

Thanks to Jeremy Wolfe, Dan Simons, and three anonymous reviewers for insightful comments. This work was supported by NSF Grant 0748689 and NIH Grant EYT018216. Correspondence concerning this article should be addressed to J. Haberman, Harvard University, Department of Psychology, 33 Kirkland St., Cambridge, MA 02138 (e-mail: haberman@wjh.harvard.edu).

REFERENCES

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, **19**, 392-398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, **106**, 7345-7350.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, **12**, 157-162.

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics*, **70**, 1325-1326.

Brown, V., Huey, D., & Findlay, J. M. (1997). Face detection in peripheral vision: Do faces pop out? *Perception*, **26**, 1555-1570.

Chong, S. C., Joo, S. J., Emmanouil, T.-A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, **70**, 1327-1334.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, **43**, 393-404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, **45**, 891-900.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, **37**, 3181-3192.

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, **70**, 789-794.

de Fockert, J. [W.], & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, **62**, 1716-1722.

Duncan, J., Ward, R., & Shapiro, K. L. (1994). Direct measurement of attentional dwell time in human vision. *Nature*, **369**, 313-315.

EKMAN, P., & FRIESEN, W. V. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

HABERMAN, J., HARP, T., & WHITNEY, D. (2009). Averaging facial expression over time. *Journal of Vision*, **9**(11, Art. 1), 1-13.

HABERMAN, J., & WHITNEY, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, **17**, R751-R753.

HABERMAN, J., & WHITNEY, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception & Performance*, **35**, 718-734.

HANSEN, C. H., & HANSEN, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality & Social Psychology*, **54**, 917-924.

KOENDERINK, J. J., VAN DOORN, A. J., & PONT, S. C. (2004). Light direction from shad(ow)ed random Gaussian surfaces. *Perception*, **33**, 1405-1420.

KUEHN, S. M., & JOLICŒUR, P. (1994). Impact of quality of the image, orientation, and similarity of the stimuli on visual search for faces. *Perception*, **23**, 95-122.

LUCK, S. J., & VOGEL, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, **390**, 279-281.

MORGAN, M. [J.], CHUBB, C., & SOLOMON, J. A. (2008). A "dipper" function for texture discrimination based on orientation variance. *Journal of Vision*, **8**(11, Art. 9), 1-8.

MORGAN, M. J., & GLENNERSTER, A. (1991). Efficiency of locating centres of dot-clusters by human observers. *Vision Research*, **31**, 2075-2083.

MYCZEK, K., & SIMONS, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, **70**, 772-788.

NOË, A., PESSOA, L., & THOMPSON, E. (2000). Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition*, **7**, 93-106.

NOTHDURFT, H.-C. (1993). Faces and facial expressions do not pop out. *Perception*, **22**, 1287-1298.

PARKES, L., LUND, J., ANGELUCCI, A., SOLOMON, J. A., & MORGAN, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, **4**, 739-744.

PESSOA, L., MCKENNA, M., GUTIERREZ, E., & UNGERLEIDER, L. G. (2002). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences*, **99**, 11458-11463.

RENSINK, R. A., O'REGAN, J. K., & CLARK, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, **8**, 368-373.

RUSSELL, J. A. (1980). A circumplex model of affect. *Journal of Personality & Social Psychology*, **39**, 1161-1178.

SIMONS, D. J., & LEVIN, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, **5**, 644-649.

SIMONS, D. J., & MYCZEK, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics*, **70**, 1335-1336.

SIMONS, D. J., NEVAREZ, G., & BOOT, W. R. (2005). Visual sensing is seeing: Why "mindsight," in hindsight, is blind. *Psychological Science*, **1**, 520-524.

SPERLING, G. (1960). The information available in brief visual presentation. *Psychological Monographs*, **74**(11, Whole No. 498).

SWEENY, T. D., GRABOWECKY, M., PALLER, K. A., & SUZUKI, S. (2009). Within-hemifield perceptual averaging of facial expressions predicted by neural averaging. *Journal of Vision*, **9**(3, Art. 2), 1-11.

TANAKA, J. W., & FARAH, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, **46A**, 225-245.

TONG, F., & NAKAYAMA, K. (1999). Robust representations for faces: Evidence from visual search. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 1016-1035.

WATAMANIUK, S. N. J. (1993). Ideal observer for discrimination of the global direction of dynamic random-dot stimuli. *Journal of the Optical Society of America A*, **10**, 16-28.

WATAMANIUK, S. N. J., & DUCHON, A. (1992). The human visual system averages speed information. *Vision Research*, **32**, 931-941.

WATAMANIUK, S. N. J., MCKEE, S. P., & GRZYWACZ, N. M. (1994). Detecting a trajectory embedded in random-direction motion noise. *Vision Research*, **35**, 65-77.

WOJCIULIK, E., KANWISHER, N., & DRIVER, J. (1998). Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *Journal of Neurophysiology*, **79**, 1574-1578.

WOLFE, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences*, **7**, 70-76.

YOUNG, A. W., HELLAWELL, D., & HAY, D. C. (1987). Configurational information in face perception. *Perception*, **16**, 747-759.

NOTES

1. The width of the region of comparison was determined by group performance in a discrimination task in which observers adjusted a test face to match a set containing identical faces. The width of the curve at three fourths of the maximum was 11 emotional units, and the region of comparison was centered at ± 60 (outlier vs. antioutlier regions).

2. Because of unusually poor baseline performance in the no-outlier (catch) trials, a performance ceiling was imposed for Subject N.W.H. We determined this ceiling by calculating the MS_e of randomly generated responses from the entire range of possible responses (uniform distribution).

(Manuscript received July 13, 2009;
revision accepted for publication May 25, 2010.)